Mu Alpha Theta National Convention: Denver, 2001
Statistics Topic Test Solutions– Open Division

For solutions to problems 6, 13, 18, 24, 27, 32, 35, 36, 37, 38, 39, and 40, please turn to the end of the document.

1. range (data) = max (data) − min (data)
$$= 10 - -5$$
$$= 15$$

2. mode (data) = the number with the highest frequency in the data set
= 8 before and after the other friend joins, so

change in mode = 8 − 8 = 0

3. $\text{Mark}_{BA} = \frac{10}{30} = \frac{1}{3} \approx .\overline{3}$

$\text{Jose}_{BA} = \frac{2}{5} = .4$

$\text{Susan}_{BA} = \frac{7}{7} = 1$

$\text{Jeff}_{BA} = \frac{7}{20} = .35$

$\boxed{\text{Lowest B.A.} = \text{Mark}}$

4. $\dfrac{\text{two innings}}{\text{five earned runs}}$ so we expect, extrapolating,

$$\frac{2 \text{ innings}}{5 \text{ earned runs}} \approx \frac{9 \text{ innings}}{X \text{ earned runs}}$$

$$\boxed{X = \text{expected earned runs} = 22.5}$$

5. $.6 \cdot (0) + .2(1) + .2(2) = \boxed{.6 \text{ fish}}$

(Expectation of a random variable $E(X) = \Sigma X \; P(X = X)$)

6. $\cancel{P(\text{movie}) = P(\text{movie} \mid \text{rain})\, P(\text{rain}) + P(\text{movie} \mid \text{sun})\, P(\text{sun})}$

$\cancel{= (.3)(\text{[illegible]}) + (.1)(.3)}$

$\cancel{= .24 \text{ or } \boxed{24\% \text{ chance}}}$

7. $E(5 \text{ slot machine plays}) = 5E(\text{slot machine play})$

(Expectation)

$= 5 \cdot (.98)$

$= \$4.90$

8. Expected return is 98% means

E.R. = (.98) X

X = money bet

Here, X = 5, so E.R. given X = 5 is .98 · 5

$= \$4.90$

9. $X_1 = \$5$

$E.R_1$ (after one try) = \$4.90

So

Expected Second Bet = \$4.90

$E.R_2$ (on expected second bet) = (.98)(4.90)

and

$E.R_3$ (after third bet w/ $E.R_2$) = (.98)(.98)(4.90)

$= \$4.71$

10. Mean time = $\frac{1}{4}\sum_{i=1}^{4} X_i$ = 4.5 seconds

fastest time = 4.0 seconds.

$\boxed{\text{Answer} = 4.5 - 4.0 = .5 \text{ seconds}}$

11. If the data is distributed symmetrically about the mean, then the data pts. on each side of the mean cancel each other as you calculate the median so that only the mean remains in the middle, and is hence also the median.

On the other hand, it is easy to create counterexamples to answers A, B, D.

12. The median is generally a more robust estimator of a data set than the mean, and is usually less affected by outliers than the mean is.

For <u>most</u> example data sets that satisfy the problem statement the mean <u>will</u> change more than the median. But not always. Consider the set {15 arrivals at 7:45 AM, 15 arrivals at 7:50 AM}.

The original mean = original median = 7:47 and 30 sec.

Add the new data pt of 9 AM. The Δ median = $2\frac{1}{2}$ min, the Δ mean = 2.3 min. So it's <u>not possible to tell</u> from the problem information which changes the most.

13. the problem says, choose W to:

$$\underset{W}{\arg\min}\ \text{abs}(W-1) + \text{abs}(W-4) + \dots$$
$$\dots\ \text{abs}(W-8) + \text{abs}(W-17)$$

the answer is the mathematical definition of the median, and thus any place between the 4th and 8th block will give the same drive.

14. The problem states a correlation:

squirrels ↑
NYSE ↑

A bothersome issue in statistics is that people assume causation when there is only correlation.

Thus the answer is B.

Furthermore, there is nothing in the ~~experiment~~ problem →

14 continued : that states an experiment, nor a goal, so (A) is not appropriate. (C) is not a good choice because the populations or data for NYSE and squirrels are large. (D) simply doesn't have anything to do with the information given.

15. It's not possible to tell from the given info. For instance, the information given is consistent with either of the following (mutually exclusive) statements :

i) No students in sports smoked cigarettes

ii) All students who smoked were in sports.

16. A - D <u>are</u> possible answers.

17. Nothing can be concluded because we don't know what proportion of the test takers were cyan or magenta.

18. You seek : $P(2 \text{ kids} \mid suv)$

by Bayes Law: $$P(2 \text{ kids} \mid suv) = \frac{P(suv \mid 2 \text{ kids})\, P(2 \text{ kids})}{P(suv)}$$

where

$$P(suv) = P(suv \mid 2 \text{ kids})\, P(2 \text{ kids}) + P(suv \mid 1 \text{ kid})\, P(1 \text{ kid})$$

(total cond. prob. law)

so $= (.25)(.6) + (.1)(.4) = .19$

$$P(2 \text{ kids} \mid suv) = \frac{(.25)(.6)}{.19} = \boxed{\frac{.15}{.19} \approx 79\%}$$

19. Var of $\{X\} = E(X - \bar{X})^2 = 1256$

Mu Alpha Theta National Convention: Denver, 2001

Statistics Topic Test Solutions– Open Division

20. School Avg $= \dfrac{28 \times 78\% + 30 \times 81\%}{(28+30)}$

21. std. dev $= \sqrt{Var} = \sqrt{5.76} = 2.4$

22. Top score = 94, so curve = 6.

Then highest to low A = 4 students

(their original scores were 84, 86, 88, 94).

23. Since every one's score increased by 6, the mean increases by 6.

~~24. We need : P(rolling a two before an odd #) = &~~

~~But for this event to happen, either we roll a two on the first try ($P = \frac{1}{6}$) or we get another even # and then we have another shot at rolling a two before an odd #.~~

25. (A) is an assumption and even if true it is not clear it would make the test invalid.

(B) not true because everyone thought they had real pills, so this would not invalidate

(C) this is a testing problem! The test was invalid in the statistician's eye because a crew member was more likely to not get sick and so the distribution of pills and placebos biased the outcome. ~~Also, there is~~ [illegible]

(D) this makes an assumption about orders and misses the point captured in (C) — "the test was biased".

26. The experimental variable is "smoking a cigarette" and the test group is "smokers". Thus answers "A" and "B" change the test group, which is not relevant. A test group of smokers that do not smoke a cigarette is an important control group, but another group that chews gum would be redundant and a confusion. Thus "C" is not a good answer. The "D" answer controls the experimental variable and the additional test group of smokers smoking a tobacco-less cigarette provides additional information about whether * any measured blood level change is due to the tobacco or to the "smoking".

27. Answers "C" and "D" result in arbitrarily long max. wait times. The arrangement "A" is preferable to "B" because one very slow customer will not force all the customers behind him/her to wait – they go to the next available cashier. A more mathematical analysis of queueing problems can be found in a Stochastic Processes book.

28.

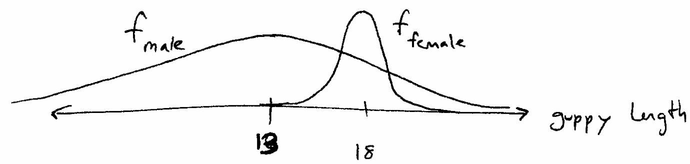

Figure above shows the high-variance male pdf and narrow-variance female pdf.

Solving for $p($guppy is female | length = 16 mm$)$

and

$p($guppy is male | length = 16mm$)$

You will find that the probability it is male is higher (compute the prob. under each gaussian).

29. $p = 3/4$, $n = 7$, The best model for X = # of successes in n trials w/ probability of success = p is the binomial distribution because p and n are moderate.

30. $p = .000023$ $n = (100)(365)$

The best model for Y = # of successes in n trials w/ probability of success = p is the Poisson distribution because p is very small and n is very large.

31. In general, a gaussian can be normalized to $N(0,1)$ by taking the original random variable X and:

$$\left(\frac{X-\mu}{\sigma}\right) \sim N(0,1).$$

By the central limit theorem, sums of iid r.v.'s go towards gaussians, and so sum of six rolls of a die is starting to look gaussian and in answer (c) it is normalized approximately to $N(0,1)$. None of the other answers come close.

32. See solution at end of document.

33. The best way to see this is to compare A-D for some different data sets to the median.

34. $S \sim N\left(E(W_1 + W_2 + W_3 + W_4),\ var(W_1 + W_2 + W_3 + W_4)\right)$

where E = expectation

W_i = ith week salary

$E(W_1 + W_2 + W_3 + W_4) = 4E(W_i) = 4 \cdot 10 = \$40$

$var(W_1 + W_2 + W_3 + W_4) = 4\ var(W_i) = \$16.$

Sum of gaussians is gaussian, thus $\boxed{S \sim N(40, 16)}$

Mu Alpha Theta National Convention: Denver, 2001

Statistics Topic Test Solutions– Open Division

6. 87th percentile corresponds to a z-score of 1.13 (1.12 leaves you just short of .87). Use the equation: $z = \frac{x-\mu}{\sigma}$. $1.13 = \frac{x-100}{16}$ x=118.08. 118 is the closest integer answer (D).

13. The mean is $\frac{4.3+7.15}{2} = 5.725$, so the margin of error is 1.425. The margin of error for a 99% confidence interval is found by $z_{\alpha/2} * S.E.$ $2.575\frac{s}{\sqrt{85}} = 1.425$. Solving gives $s = 5.1$ (C)

18. We do not need to know that the parent population is normal, due to the sample size. Use the equation $z = \frac{\bar{x}-\mu_\circ}{\frac{s}{\sqrt{n}}}$. $z = \frac{260-250}{\frac{50}{\sqrt{100}}} = 2$ $P(z > 2) = .0228$ (A)

24. Margin of error for a confidence interval is $z_{\alpha/2} * S.E.$ The standard error for this problem is $\sqrt{\frac{pq}{n}}$ solving the equation $.01 = 1.96\sqrt{\frac{.25(.75)}{n}}$ gives $n = 7203$ (B).

27. The appropriate test is the Chi-squared test of independence. Expected values are found by multiplying the row margin by the column margin and dividing by the total. Chi-squared is the sum of (observed minus expected) squared, divided by the expected. Chi-squared for this test was 3.04. there are 2 degrees of freedom, making the p-value greater than .1 (A).

32. This is paired date, so a new statistic will be computed: after minus before for each subject. We expect the cholesterol to be lowered, so we are testing whether the mean of the new statistic is less than zero. Use the equation: $t = \frac{\bar{x}-\mu_\circ}{\frac{s}{\sqrt{n}}}$

Substituting gives $t = \frac{-8.625-0}{\frac{15.528}{\sqrt{8}}} = 1.571$ (ignore the sign). There are 7 degrees of freedom. In the t table for 8 d.f., 1.571 shows p to be greater than .05 and less than .1 (B)

35. Only III is correct – look up the definition of least squares regression!

36. I and III are ridiculous. r^2, the coefficient of determination, tells what portion of the variation in the data is explained by the model.

37. Quadrupling the sample size cuts the standard error in half. The Central Limit theorem says that as n becomes large, the standard error (standard deviation of the sampling distribution) becomes smaller. A type II error occurs when you fail to reject a false null hypothesis. All are false. (E)

38. The appropriate test is a z-test for difference of proportions. $z = \frac{p_1 - p_2}{S.E.}$.

$S.E. = \sqrt{\frac{p_1 q_1}{n} + \frac{p_2 q_2}{n_2}}$ Substituting gives $z = 1.93$. This is a two tail test, so the p-value is 2*.0268=.0536 (B)

39. Since there are 5 groups, the d.f. treatment is 4. Looking in the α=.05 F table, 4 numerator degrees of freedom, we find 2.71 corresponds to 28 degrees of freedom (error). Since d.f. error is N minus k, there must be 33 samples in the entire study (D)

40. Interquartile range is $Q_3 - Q_1$. This is represented by the width of the box. Only A has interquartile range = 30.